Lab #3: Correlation Analysis

A. Objectives

1. Compute and interpret Pearson correlation between numeric variables.
2. Use cross-tabulation to examine relationships between categorical variables.
3. Assess numeric–categorical relationships using plots and tests.
4. Present clear graphs and interpretations for all variables.

B. Theory

Pearson Correlation:

- A measure of the **linear relationship** between two numeric variables (ranges from –1 to +1).
- **Data type required: Numeric–numeric** (e.g., income vs education years).
- **Why used:** To check whether increasing one variable tends to increase/decrease the other, and how strongly they are related.

$$r = \Sigma(x-\bar{x})(y-\bar{y}) / \text{sqrt}(\Sigma(x-\bar{x})^2 \Sigma(y-\bar{y})^2)$$

Cross -Tabulation:

- A table showing **frequency counts** of two categorical variables.
- **Data type required: Categorical–categorical** (e.g., gender × internet access).
- **Why used:** To observe patterns, distributions, or associations between categories before performing statistical tests.

Chi-Square Test:

- A statistical test to check whether **two categorical variables are independent**.
- **Data type required: Categorical–categorical** with a crosstab.
- **Why used:** To determine whether differences in the crosstab are due to **real association** or just **random chance**.

$$\chi^2 = \Sigma\ (O-E)^2 / E$$

t-test:

- t-test is a statistical test used to determine if there is a significant difference between the means of two groups
- A test comparing the **mean of a numeric variable across two groups**.

- **Data type required:**
  - One **numeric** variable
  - One **categorical** variable with **2 categories** (e.g., private vs public school)
- **Why used:** To see if the two groups differ **significantly** (e.g., do private-school students score higher?).

$$t = (\bar{x}1 - \bar{x}2) / \text{sqrt}(s1^2/n1 + s2^2/n2)$$

ANOVA (Analysis of Variance):

- A test comparing the **means of a numeric variable across 3 or more groups**.
- **Data type required:**
  - One **numeric** variable
  - One **categorical** variable with **3+ categories** (e.g., math score across 7 provinces)
- **Why used:** To check whether **at least one group** differs significantly from others.

$$F = MS_{between} / MS_{within}$$

C. Dataset Description

Dataset: nepal_education_survey.csv

D. Observations

Students must show output and descriptions of the outputs

- **Pearson Correlation , Scatterplots & heatmap** (for at least family_income_npr vs education_years , education_years vs math_score. Also, explain about the interpretation /significance of the values/plots of these observations.)
- **Crosstabs and Chi-square tests** (for at least area × school_type, province × internet_access. Also, explain about the interpretation /significance of the values/plots of these observations.)

- **Box Plot** (for at least Math Score by School Type, Family Income by Area. Also, explain about the interpretation /significance of the values/plots of these observations.)

- **t-tests and ANOVA** (for at least (Math): Public vs Private, (Income): Urban vs Rural. Also, explain about the interpretation/significance of the values of these observations.)

E. Conclusion

Submission Guidelines

Complete the Jupyter notebook with observations. Document each step with comments and explanations. Submit the well documented notebook file (.ipynb) to bidur(@)gces.edu.np with email

- **subject**: BECE2022 - CMP 360 – Lab#3.
- **body** must have your: Name, Class Roll Number, Lab Number and Lab Title.
- Use your **gces email** to complete the submission.

Hardcopy Submission (Individual Handwritten):

- Lab Title
- Objectives
- Theory
- Observation
- Conclusion